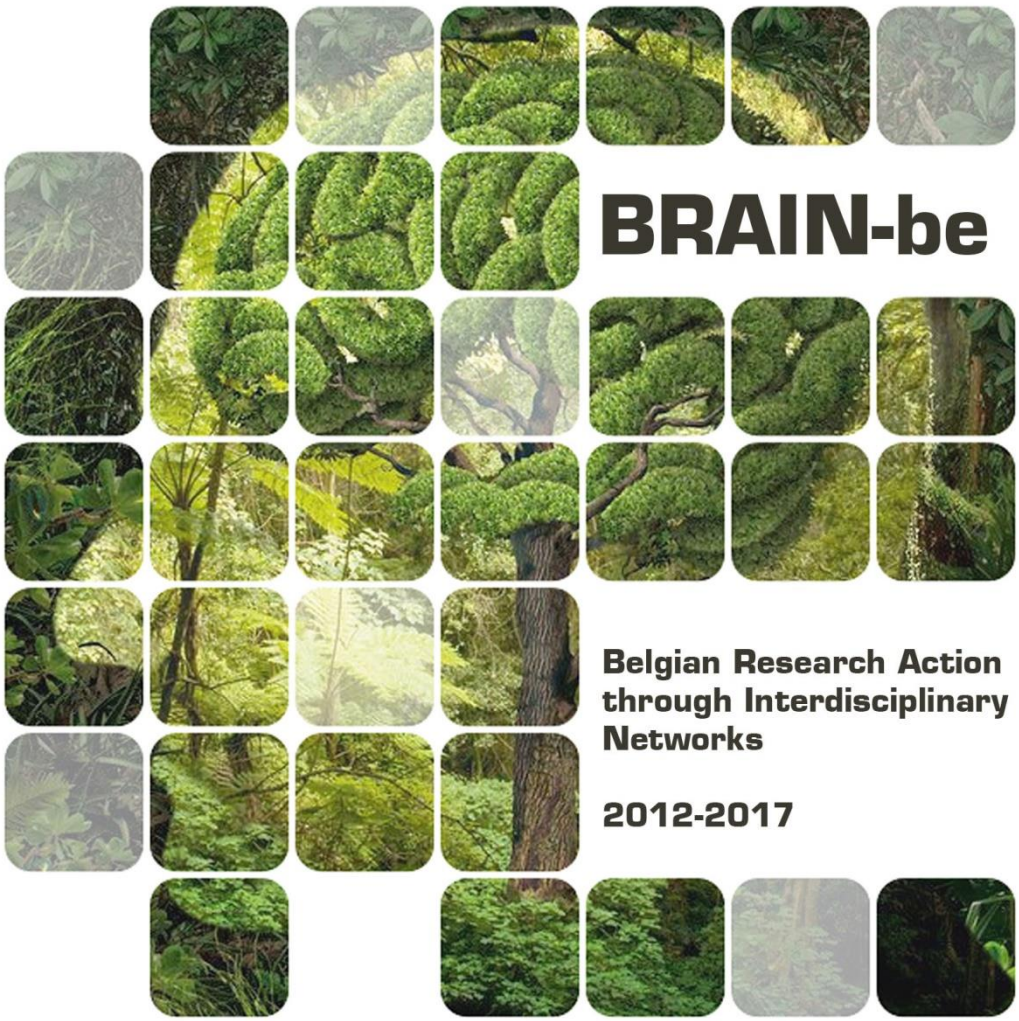
PROMISE : Preserving Online Multiple Information

Di Pretoro, Emmanuel; Geeraert, Friedel; Merchant, Peter; Michel, Alejandra

Publication date:
2020

Document Version
le PDF de l'éditeur



PROMISE

Preserving Online Multiple Information: towards a Belgian strategy

Emmanuel Di Pretoro (HE2B, URFSID) - Friedel Geeraert (KBR & State Archives) - Peter Mechant (Ghent University, MICT) - Alejandra Michel (University of Namur, CRIDS)

Axis 3: Cultural, historical and scientific heritage Axis 6: Management of collections

belspo .be

NETWORK PROJECT

PROMISE

Preserving Online Multiple Information: towards a Belgian strategy

Contract - BR/175/A3/PROMISE

FINAL REPORT

PROMOTORS:
Sophie Vandepontseele & Nadège Isbergue (KBR)
Rolande Depoortere & Sébastien Soyez (AGR)
Cécile de Terwangne & Benoit Michaux (UNamur)
Emmanuel Di Pretoro (HE2B)
Peter Mechant & Sally Chambers (UGent)

RESEARCHERS:
Emmanuel Di Pretoro, HE2B-URFSID
Friedel Geeraert, KBR & AGR
Gerald Haesendonck, UGent- IDLab
Alejandra Michel, UNamur-CRIDS
Eveline Vlassenroot, UGent-MICT

AUTHORS:
Emmanuel Di Pretoro, HE2B-URFSID
Geeraert Friedel Geeraert, KBR & AGR
Peter Mechant, UGent-MICT
Alejandra Michel, UNamur-CRIDS

Published in 2020 by the Belgian Science Policy Office
WTCIII
Simon Bolivarlaan 30 Boulevard Simon Bolivar
B-1000 Brussels
Belgium
Tel: +32 (0)2 238 34 11 - Fax: +32 (0)2 230 59 12
http://www.belspo.be
http://www.belspo.be/brain-be

Contact person: Georges Jamart
Tel: +32 (0)2 238 36 90

Di Pretoro Emmanuel, Geeraert Friedel, Mechant Peter, Michel Alejandra. ***PROMISE: Preserving Online Multiple Information: towards a Belgian strategy.*** Final Report. Brussels: Belgian Science Policy Office 2020 – 36 p. (BRAIN-be - (Belgian Research Action through Interdisciplinary Networks))

TABLE OF CONTENTS

ABSTRACT

This document is the final report of the PROMISE research project that focuses on developing a federal strategy for the preservation of the Belgian web. The web has become a central means of communication in our everyday lives, which makes it very valuable from a heritage perspective and therefore worth preserving. Within the project four main goals were addressed: 1) to identify current best practices in web archiving, 2) to define a Belgian policy for web archiving on the federal level, 3) to pilot web archiving, access and use of the pilot Belgian web archive for scientific research and 4) to make recommendations for a sustainable web archiving service for Belgium. The project ran between 2017 and 2019 is a collaboration between KBR, the State Archives (AGR), Ghent University, Namur University and the Haute-Ecole Bruxelles-Brabant. The research results cover a number of disciplines: information management, media and communication studies, digital humanities, ICT and law. The main deliverables comprise of a comprehensive report on best practices within the field of web archiving, an in-depth legal analysis of the Belgian legal framework surrounding web archiving and providing access to the collections, an analysis of user requirements, a web archiving strategy for the State Archives and KBR, a prototype for the Belgian web archive and recommendations for a sustainable web archiving service in Belgium.

1. INTRODUCTION

The web has become a central means of communication in our everyday lives, which makes it very valuable from a heritage perspective. Websites, as collections of data and documents, are therefore important materials to be archived. Today the web is also considered as a publication channel in its own right. As is the case for other publications and archives, the preservation of which is guaranteed by legal deposit legislation and the law on archives, a long-term preservation policy needs to be developed for the Belgian web. The PROMISE project was initiated to formulate an answer to the urgent question of how to address the preservation of the Belgian web for future generations, as an important part of Belgian history.

2. STATE OF THE ART AND OBJECTIVES

On an international level, many national libraries and national archives have been preserving their national web for decades. The first web archives were established in 1996 (Australia and UK) and many countries followed suit: Sweden (1997), New-Zealand (1999), USA and Czech Republic (2000), Norway (2001), France and Japan (2002), Croatia and Iceland (2004), Denmark, Latvia and South-Korea (2005), etc. (Schroeder and Brügger (2017). Except for Italy, Lithuania and Poland, Belgium is the only European country that currently does not have a web archive at the national level. Belgium has already lost 25 years of web history and if nothing is done to remedy this, researchers wishing to study the period from the 1990s onwards will be faced with a digital dark age. As Ian Milligan (2016) states: “For the most part, historians cannot write histories of the 1990s unless they use web archives: with them, military historians will have access to the voices of rank-and-file soldiers on discussion boards; political historians, to blogs, the cut and thrust of websites, electoral commentary and beyond; and of course, social and cultural historians, to the voices of the people on a scale never before possible.” Niels

Brügger, one of the most prominent researchers in the field of web archives as a resource for historical research, remarked that having a comprehensive web archive is comparable to having a tape recorder on a market square in the Middle Ages, given the diversity of voices that are present on the web that can be captured.

Web archives form a real treasure trove for research because of the diversity of sources that are available for researchers. A plethora of research fields can use these sources for their research: linguists, historians, musicologists, social and political scientists, information, media and communication scientists, etc. Web archive content can for example be used to trace the historical development of a national web over the course of several years, as is the case in Denmark, or to zoom in on a specific period in history such as the French web during the 1990s. (Brügger (2014), Agence nationale de la Recherche (s.d.)) The representation of certain movements such as the extreme right in Spain (Ben-David and Matamoros-Fernández (2016)) can also be studied as well as important national events such as terror attacks (ASAP (s.d.)). On a smaller scale, web archives can also be used to do qualitative diachronic studies of certain websites or web pages such as the evolution of a blog of a French person living in London (Huc-Hepher (2016)) or as part of the source material for cross-media studies, for example to study North-African immigration memories (Gebeil (2015)). In the future transnational studies within web archives will be made possible in order to study transnational events such as the refugee crisis, the European elections, etc. Given their size, web archives are also particularly suited for using computational methods and big data analysis such as sentiment analysis (Maynard et al. (2013)), topic modelling (Milligan (2015)), network analysis (McTavish (s.d.)) and machine learning (Elshobaky, (2019)), etc.

The lack of a Belgian web archive excludes all these research possibilities. We will never be able to comprehensively study the evolution of the early Belgian web, we will not be able to analyse the reactions on social media to the terror attacks of 2016, etc. It is therefore crucial that the Belgian web is preserved for future generations.

Setting up a Belgian web archive would also facilitate the preservation of the websites of the Belgian federal institutions, something that is, in theory, required under the Law on Archives, however does not yet happen in practice. For KBR, web archiving is a logical extension of the legal deposit. With the dawn of the digital era and the rise of online publications, the notion of a ‘publication’ has widened and has found its own place in the digital world and the world of the web. The web is considered today as a publication channel in its own right. For the State Archives and KBR a web archiving strategy would enable the creation of collections that are unique in Belgium. A number of smaller web archiving initiatives exist in Belgium (Felixarchief, Liberaal Archief, AMSAB, Ghent University Library, Archief Gent, etc.), but their collection scopes are narrow, meaning that only a very limited part of the Belgian web is currently preserved.

The PROMISE project addresses this challenge and aims to develop a federal strategy for the preservation of the Belgian web. Four goals are set for the project:

- To identify current best practices in web archiving

- To define a Belgian policy for web archiving on the federal level
- To pilot web archiving, access and use of the pilot Belgian web archive for scientific research
- To make recommendations for a sustainable web archiving service in Belgium

3. METHODOLOGY, SCIENTIFIC RESULTS AND RECOMMENDATIONS

Work package 1: Web archiving: state of the art (lead: KBR)

Task 1.1 Review of existing web-archiving projects (MICT and GhentCDH)

Task 1.2 Analysis of legal frameworks for web-archiving (CRIDS)

Task 1.3 Analysis of the implementation of legal framework for web-archiving (CRIDS, AGR and KBR)

Task 1.4 Analysis of existing technical solutions for web-archiving (MICT and URF-SID)

During the first phase of the project, for the identification of current best practices in the field of web archiving or the state of the art, a mix of different approaches were used. First a secondary approach (desk research) was taken. This involved summarising, collating and/or synthesising documentation related to existing web archiving projects in order to gain a general overview of web archiving initiatives and their scopes. Secondly, a number of web archiving initiatives were selected to be studied in more detail. The selection of these initiatives was based on the following characteristics that were deemed valuable within a Belgian context:

- Established web archiving initiatives with ample experience
- Web archiving initiatives in countries where both the National Library and the National Archives are involved in web archiving (as the PROMISE project is a collaboration between the Belgian Royal Library and State Archives, useful lessons could be drawn from countries where both institutions engage in web archiving)
- Web archiving initiatives in countries with multiple official languages
- Web archiving initiatives in countries of different sizes
- Combination of web archiving initiatives relying on external service providers and initiatives that manage all aspects of the process in-house.

The following web archiving initiatives were selected to be studied in further detail with regards to selection and access policies and legal and technical framework:

1. **The Netherlands**
    a. National Archive of The Netherlands
    b. National Library of The Netherlands
2. **France**
    a. National Library of France
    b. Institut national de l'Audiovisuel

3. **Luxembourg: National Library**
4. **United Kingdom**
    a. British Library
    b. UK National Archives
5. **Denmark: Royal Library**
6. **Portugal: Arquivo.pt**
7. **Ireland: National Library**
8. **Canada**
    a. Library and Archives Canada
    b. National Library and Archives Québec
9. **Switzerland: National Library**

Interviews were conducted with representatives of these institutions, although in some cases the representatives that were contacted preferred to reply to the questions in writing, as was the case for Institut national de l'Audiovisuel (INA) in France and National Library and Archives of Québec. All the other interlocutors were interviewed through face-to-face meetings or conference calls. The interviews were semi-structured, using pre-defined as well as open questions. All interviewees were sent a list of questions prior to the interview. Some participants already provided written replies to some of these questions, in which case the interview consisted mainly of follow-up questions.

For the analysis of the legal frameworks for web archiving, the Belgian legal framework was analysed first: on the one hand the legal deposit and on the other hand the framework related to archives and public records. In this context, the legal instruments assigning the missions to KBR and AGR were also examined.

After examining the Belgian law in this area (Library and Archives), the existing legal situations in several European countries as well as outside Europe were analysed. The foreign legal framework surrounding the activities of National Libraries and National Archives were studied. To do this, three sources were chosen: national legal instruments, interviews and legal doctrine. It was noticed that these different countries approach web archiving on a very different basis: while some have amended their law on the legal deposit to widen it to include web legal deposit, others have no legal deposit and rely either on the general mandate to preserve national cultural heritage or on the opt-out approach to undertake web archiving. This study of foreign legal frameworks allowed us to gather information on the functioning of these foreign institutions in various legal fields: respective missions of national cultural institutions, legal deposit, archives, copyright (both on selection and access levels), protection of personal data (especially the right to be forgotten and the right to rectification), national scope of competence/jurisdiction for web archiving, illegal contents, authenticity and integrity of web archives, etc.

The major result of WP1 from the legal point of view was the identification of a number of criteria to define and delimit the national scope of competence/jurisdiction for web archiving in the various

countries studied. It should be noted that the criteria used abroad are varied and that these criteria were analysed in the second work package to see which are the most relevant for the Belgian situation.

The study of foreign legal frameworks was concluded by making a general comparison from a legal point of view between all these legal aspects which were very useful for the realisation of the second work package, namely the definition of a Belgian policy for web archiving. This comparison also contains many interesting elements in foreign web legal deposit laws that inspired the recommendations for the amendment of the Belgian legal deposit law.

A third step within this first research phase, focusing on the state of the art of web archiving, encompassed further validation and synthesis in order to reach an overarching view on the selected web archiving initiatives. It drew comparisons between the different initiatives with regards to selection and access policies, technical infrastructure and legal frameworks and in doing so, distilled the aspects that required further reflection or from which inspiration could be drawn in the later stages of the PROMISE project.

Work package 2: Definition of a Belgian policy for web archiving (lead: AGR)

The purpose of this work package was to define a Belgian policy for web archiving and was divided into 5 tasks.

Task 2.1 Analysis, typology and classification of 'Belgian web information' (URF-SID and MICT)

In order to make an analysis, typology and classification of Belgian web information, it was first necessary to know what 'Belgian web information' is and how to obtain it. A number of techniques were implemented:

- to gather as many active .be, .vlaanderen, .gent and .brussels websites as possible;
- to obtain geographical information on hosts serving websites; this allows to know which sites - not only .be, .brussels, .gent and .vlaanderen - are hosted in Belgium.
- to crawl a representative sample of these websites selected by AGR and KBR.

This information has been gathered using a combination of publicly available data from Wikipedia and Common Crawl (Common Crawl (2019)), and websites crawled by the project team, as well as a sample of URLs sent by DNS Belgium (DNS Belgium (2019)). Geo-IP localisation was used to identify websites hosted in Belgium that are outside of the .gent, .vlaanderen, .brussels and .be domains.

The various crawls carried out during this period made it possible to collect figures allowing for a more precise idea of what a Belgian website was. It was thus possible to have some descriptive statistics about an average Belgian website: total size, number of pages, number of images, number of image sizes, etc.

Desk research was undertaken to find methodologies in the field of classification / clustering and metadata extraction that can be applied to the data. Concerning the classification and clustering of

web pages, several methodologies were identified. However, given the nature of the methodologies, namely supervised machine learning, it was not possible to quickly test these methodologies in the time available within the project. For metadata extraction, a prototype has been developed using Apache Tika. This prototype was a proof of concept and was not used in the project. However, it is interesting to note that the indexing tool used to feed WARCLight, the discovery tools used to perform full-text searches in web archive collections, also uses Apache Tika to extract certain metadata such as title, author, etc. This indexing tool is called webarchive-discovery.

Finally, several prototypes have been developed using different technologies, namely FIDO and DROID, to determine possible archiving problems with the resources contained in the WARC files. These prototypes were mainly proofs of concept, and were not used in the prototype.

Task 2.2 Analysis of the legal framework for eBelgian web information (Lead: CRIDS, AGR and KBR)

This task consisted of analysing and formulating answers to a number of questions:

1) What can be considered as the "Belgian web" and what constitutes "Belgian online information"?

Based on the information gathered during the first work package, an analysis was done with regards to how 'the national web' as a concept had been delimited in foreign legislation on web legal deposit and on archives. An exhaustive list of criteria drawn from legal deposit legislation was established. These criteria were critically analysed and annotated. The legislation surrounding archives showed that the web archiving activities can take place mostly because the notion of what constitutes an 'archive' or 'record' is very broadly defined. In addition, wherever possible, a link has been made between the legal provisions and the way in which these legal elements are translated on the operational level. To conclude this question, several recommendations were made linking the legal and operational aspects for the definition of the "Belgian web".

2) What are the roles and responsibilities of the Royal Library and the State Archives of Belgium?

This question was answered based on the study and analysis of the legal missions and legislation surrounding the KBR and AGR.

3) Which elements in foreign legal frameworks on web archiving are interesting for the Belgian situation?

The results of the analysis of foreign legal frameworks in Work Package 1 allowed for the creation of a detailed list of interesting elements that can be taken into account when further developing the Belgian web legal deposit legislation and legislation on archives to include web archiving.

4) What is the probative value of online information?

The probative value of online information is strongly linked to the concepts of *authenticity* (a document is what it purports to be) and *integrity* (a record is complete and unaltered). Both concepts

were analysed from a legal - either at the EU level with the eIDAS Regulation, or at the Belgian level with the Digital Act - and from an operational perspective. It was shown that, given the complex nature of websites as objects to be archived, it is difficult to guarantee the integrity of captured web content compared to the web content on the live web. What can be guaranteed however is that the content has not changed after it has been captured and ingested into the web archive. Authenticity is strongly linked with metadata management and documentation. Sufficient attention should be paid to these aspects when defining procedures.

A summary document was also produced with the results of this task, including the legal basis and competences of each institution, the reminder of copyright rules and reasoning schemes as well as legal recommendations related to copyright. The findings of this legal analysis were translated into operational guidelines by means of decisional flowcharts that show what the State Archives and KBR are legally allowed to do or not.

Task 2.3 Proposals for a global Belgian web archiving strategy (AGR, KBR and CRIDS) and Task 2.4 Selection criteria for appraisal methods (AGR and KBR)

On the operational level, a proposal for a global Belgian web archiving strategy and selection criteria was developed. The selection criteria that were outlined were based on the analysis of the Belgian legal framework and the selection policies of the web archiving institutions that were studied during the first research phase. The web archiving strategy is based on the OAIS model or the Open Archival Information System which is also an ISO standard ISO 14721:2012 (ISO (2018)). The strategy describes in detail the vision of KBR and the State Archives on a shared web archive at the federal level. The following phases are included: *Ingest, Data Management, Archival Storage, Access, Preservation Planning and Administration*. Two additional phases were added to align the model with the reality of web archiving: *selection* and *capture*.

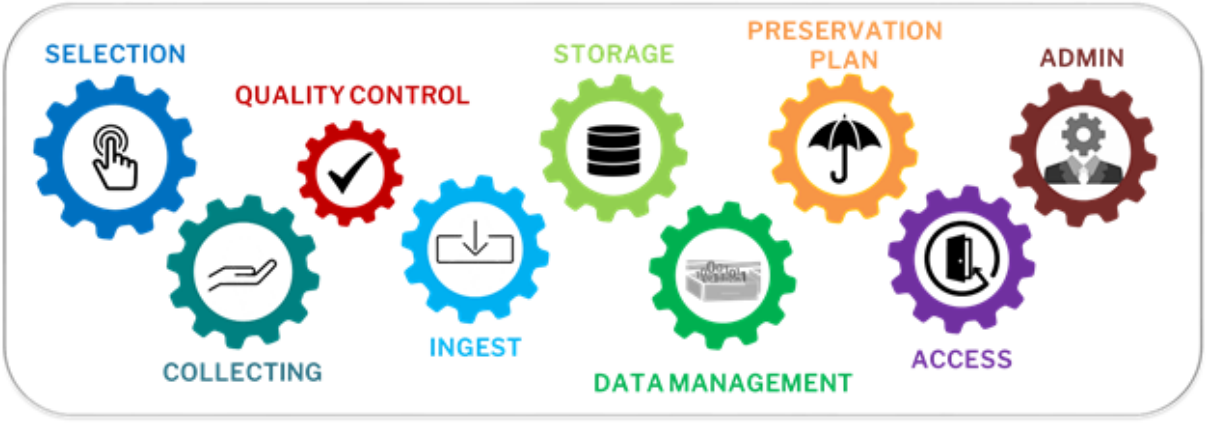


Figure 1: the different steps in the web archiving process described in the strategy

On a methodological level, the strategy is based on the research results of the first research phase of the project, namely the state of the art. The legal considerations and policies with regards to selection

and access that were identified abroad were also taken into account. Furthermore, the results of the analysis of user requirements (see Task 2.5) and the analysis of the Belgian legal framework (see Task 2.2) informed the development of the strategy.

In addition to this detailed description of the different phases in the web archiving workflow, a detailed cost analysis was also undertaken in order to work out specific scenarios that could be used by the Management of KBR and the State Archives to decide which web archiving strategy to choose in the future. Different selection policies form the basis for these different scenarios. The most elaborate selection policy would be to combine a broad crawl with selective crawls. In this scenario, the broad crawl would comprise capturing 100% of the Belgian web at a superficial level. In contrast, the websites included in the selective collections that are based on the legal missions of KBR and the State Archives and on their existing collections would be completely captured. The most limited selection policy would be to focus only on selective crawls for the State Archives and KBR.

A list of tasks divided by phase (selection, capture, ingest, data management, etc.) was drafted. Function titles were assigned to each task as well as estimations regarding the necessary investment of time each task would require. The number of hours assigned were thereafter multiplied by the average hourly wage of each function profile in order to calculate the total cost of human resources per scenario. The cost for technical infrastructure was also calculated based on 5-year projections with an estimated 10% increase in data size each year.

The resulting strategic note was presented to the Management of both institutions in September and October 2019 and will serve as a basis for the implementation of a operational web archive.

Task 2.5 Analysis of user requirements for a Belgian web archive (MICT and GhentCDH)

In order to take into account the changes and challenges imposed by a common preservation and archiving strategy for the web, especially in terms of consultation of the archive, an online survey was developed that polled for the user requirements of a Belgian web archive. This survey was jointly agreed upon and developed by the project partners during February and March 2018. Existing literature (e.g. Costa & Silva (2010;2011), Costea (2018), Weber (2016)) as well as insights from informal talks with information professionals (e.g. KBR, AGR, VIAA, etc.) were used to create the survey questions. Survey questions were collaboratively refined and rephrased during an iterative process (mainly via online tools but also via F2F-meetings). Various answer-paths were developed as well as a Data Protection Notice. The survey was made available in 3 languages (English, French and Dutch) using Qualtrics and was aimed at 3 target groups:

- People working as a researcher, academic, student or in any way involved in research activities.
- People working at an archive, library, heritage or government institution.
- The general public (people active in another field/domain than the fields described above.)

The survey was launched at the beginning of April 2018 and was closed two months later. Project partners actively engaged to distribute the survey URL (e.g. via newsletters, websites, etc.) and a

dedicated effort was made to send personal survey invitations to relevant researchers and research groups. The survey data has been cleaned and general results are available via an interactive Tableau dashboard at:

https://public.tableau.com/profile/eveline.vlassenroot#!/vizhome/PReservingOnlineMultipleInformationtowardsaBelgianstrategy/PReservingOnlineMultipleINformationtowardsaBelgianstrategy.

A more in-depth analysis of this data was presented in *D2.5: The use of web archives: current state-of- play and requirements* and will be presented at the 2020 IIPC Web Archiving Conference (WAC).

Work package 3: Piloting web archiving in Belgium (URF-SID, MICT and GhentCDH)

The purpose of this work package was to set up a pilot Belgian web archive including selection of content and tools, harvesting the web content, piloting access and evaluating the pilot Belgian web archive.

Task 3.1: Selection of web content for the pilot Belgian web-archive (GhentCDH, KBR and AGR)

The common strategy and the selection criteria that were defined in task 2.3 and task 2.4 served as a basis for the selection of web content to be included in the pilot web archive. The purpose was to simulate a broad crawl and create selective thematic collections.

For the broad crawl, a random sample of 10,000 and 100,000 websites was drawn from the list of Belgian URLs established in task 2.1. The timing of the creation of the seed lists, e.g. lists of pertinent URLs to be included in the web archive for the selective thematic collections, coincided with a change in Library Management System at KBR, meaning that a number of cataloguers had no means to catalogue for a few weeks in the summer of 2018. Some of them volunteered to help create lists of pertinent URLs. They received theoretical and practical training and worked on specific thematic collections linked to the core collections of the library. The table underneath shows the collections on which the cataloguers worked. The complete seed list comprised 928 websites, 1,416 web pages and 37 sections of websites.

Table I: Overview of thematic collections created for KBR

Collections linked to the department of contemporary collections	**Collections linked to the department of heritage collections**	**Transversal collections**
Literary blogs	Restauration and conservation in Belgium	Belgian literature throughout the ages
Belgian editors	Music in Belgium	Belgian library sector

E-magazines		Belgian education system
Online fan fiction		Online representation of minorities in Belgium
History of the Belgian comic books		
Youth literature		
Literary prizes		

At the State Archives the seed list mainly focused on the following categories:

1. Websites of federal institutions: executive, judicial and legislative powers
2. Websites of ministerial cabinets, ministers/secretaries of state
3. Websites of other public organisations that have a link with the federal level: trade associations, trade unions, federations, mutual insurance companies political parties, public interest organisations, the monarchy, etc.
4. Websites of the provinces, the regions and the communities

The complete seed list of the State Archives counts 645 websites.

The creation of descriptive metadata was an important part of this task for both institutions. The descriptive metadata are based on a report by the OCLC Research Library Partnership Web Archiving Metadata Working Group (Dooley and Bowers (2018)). The study suggests a set of 14 metadata elements and includes a mapping to MARC21 and EAD which are the metadata standards used in KBR and the State Archives respectively. Table II provides an overview of this set of metadata. As both institutions wish to make their collections available via a shared access platform, the metadata set suggested by the OCLC is a useful model. As the module for the creation of descriptive metadata was still in development at that time, a template was created in Excel to make the creation of metadata as uniform as possible between both institutions. It is important to note that these metadata elements were only created for the selective thematic collections. For the broad crawl only the metadata elements that could be automatically retrieved from the websites (URL, language, title) were preserved. The sheer scale of the broad crawl would make it too time-consuming to manually create the metadata for each website.

Table II: Overview of descriptive metadata created for the thematic collections (Dooley and Bowers (2018))

URL	Internet address for an archived website or collection
Title	The name by which an archived website or collection is known
Creator	Organisation or person **principally** responsible for creating the intellectual content of an archived website or collection
Contributor	Organisation or person **secondarily** responsible for the content of an archived website or collection
Language	The language(s) of the archived content, including visual and audio resources with language components
Collector	Organisation responsible for curation and stewardship of an archived website or collection
Date	A single date or span of dates associated with an event in the lifecycle of an archived website or collection
Subject	Primary topic(s) describing the content of an archived website or collection
Genre/Form	A term specifying the type of content in an archived website or collection
Relation	Used to express part/whole relationships between a single archived website and any collection to which it belongs
Description	One or more notes explaining the content, context and other aspects of an archived website or collection
Extent	An indication of the size of an archived website or collection
Rights	Statements of legal rights and permissions granted by intellectual property law or other legal agreements
Source of description	Information about the gathering or creation of the metadata itself, such as sources of data or the data on which source data was obtained

The descriptive metadata created for the collections of KBR was first created using the OCLC metadata set. This was then mapped to MARC21 by means of a csv file and was successfully ingested into Syracuse, the catalogue at KBR.

Task 3.2: Selection of tools/collecting model for the pilot Belgian web-archive (URF-SID)

On the technical level it was initially the intention to create prototypes for three tasks: the selection of websites, capturing these websites and providing access to these collections. During the project new needs came to light and a prototype for semi-automatic quality assessment was also created as well as derivative files to facilitate the use of the collections. This section will discuss the tool for the selection of websites, the prototype for semi-automatic quality assessment and the creation of derivative files. The other elements are discussed in tasks 3.3 and 3.4.

The selection module was developed in-house based on Python, Django and PostgreSQL. The intention of this module is to allow people working on seed lists for selective thematic collections to create the descriptive metadata based on the OCLC model described above and feed them directly into the catalogues of both institutions. At a later stage, and with additional development, this module could also be used to organise the crawling and the quality control.

Regarding the semi-automatic quality analysis three important aspects related to the quality of a web archive were analysed within the project: visual correspondence of the archived version compared to the original live version, interactional correspondence or how well the interaction corresponds between both versions and completeness or how complete the archived version is compared to the original live version. For the visual correspondence two parameters were used: structural similarity and the visual quality indicator. The first is useful to spot small changes between two images, but is less useful when it comes to changes in colour. The visual quality indicator on the other hand is better suited to spot changes in colour but less suited to spot other small changes between two images. By combining both measures it can be determined how good the visual quality of an archived web page is. A prototype was developed that automatically measured the visual quality of archived web pages.

The interactional correspondence measures to what degree a user can interact with the archived website in the same way as with the original website. When a user clicks on a link on a web page for example, this results in a series of browser requests to the server (fetching the HTML documents, stylesheets, images, scripts, etc.). It can be determined how many of these requests are also successful in the web archive. Furthermore, these requests can be weighted by their importance. A prototype was created to measure the interactional correspondence. First an index is made of the archived data in order to facilitate fast look-ups. Then the web page is crawled from the web archive (not from the live web) filtering out adverts because these are not useful for user interaction. Then the importance of each element of the website is determined. HTML for example is considered to be more important than images and images as more important than fonts. Other aspects are the stylesheets and the size and position of the images on the web page. Based on this analysis the interactional correspondence can be automatically calculated.

The completeness measures the degree to which a web archive contains all the resources available on the original websites. To calculate this completeness, it is necessary to browse the live version of the website, and to check for each identified resource that it appears in the archive. A prototype has been developed in Python that calculates the cosine similarity between the resources present in the web archive, based on the CDXJ index, and those identified on the live version by a crawler. The CDXJ index is similar to the CDX format, but it simplifies the fields used to describe each resource in a web archive and it appends a JSON block at the end of each description. That JSON block can be used to add any kind of information to each record in the index. The prototype is a work in progress as it only extracts links from the live site from HTML and JavaScript files. It also does not take into account images and PDF files so that it does not slow down the calculation of completeness too significantly. However, it would be quite simple to solve the link extraction problem using a headless browser, and to configure or change the crawler used, namely Scrapy, so that it takes into account all kinds of resources.

Another element that was developed within the project was the possibility to create derivative files. These files facilitate gaining insight into a web archive as a whole. There are a number of derivative data formats that can be created based on WARC files. Within PROMISE, the Archives Unleashed Toolkit (Archives Unleashed Cloud (2019)) was used. The Archives Unleashed Toolkit is an opensource tool that allows to create derivative data formats from WARC files. These comprise a *domain file* that lists how many times each domain appears in the web archive, a *text file* that contains the plain text of all pages and documents found in the archive and finally a *raw network file* that allows to visually represent the connections between web pages by means of visualisation software such as Gephi (Gephi (2019)). Gephi can, for example, be used to see how the websites are connected via hyperlinks or which websites have a lot of incoming links and can therefore be deemed more important than others. Another functionality was developed within the PROMISE project to enable data extraction of certain domains or web pages so that 'sub-archives' can be created as a new archive. Given the smaller size of these 'sub-archives', it is easier to do research on them than on the entire archive. This was established by building an index to enable fast look-up, then crawling the websites that are considered interesting from the web archive (not the live web) and putting the output in a new sub-archive.

Task 3.3: Harvesting the web-content for the Belgian pilot web-archive (URF-SID)

For the capture of the websites the Heritrix tool was used (Internet Archive (2019a)). This is an opensource tool that works as a big spider. It starts from a list of URLs, captures the content on these web pages and checks which other links are included in the web pages. This process is repeated over and over again until an entire website is captured. The content is saved in the WARC file format that is an ISO standard and is widely used in the field of web archiving (ISO (2017)). The advantage of Heritrix is that it is a tool that is very well equipped to do broad crawls because of the speed with which it crawls. Heritrix can also be easily configured (for example which domains not to capture, how often you would like to recapture a certain website, how many clicks deep websites should be captured, etc.) . It is also very stable software. The main disadvantage of Heritrix is that it cannot cope very well with complex websites that contain a lot of social media content or Javascript. Other tools such as Browsertrix (Webrecorder (2019a)) and Brozzler (Internet Archive (2019b)) were tested. These

tools work in a similar way to Heritrix but use a real browser without a graphical interface. The tests resulted in high-quality crawls. However, these tools are considerably slower than Heritrix and not as stable since they are rather experimental. The decision was therefore made to use Heritrix during the pilot.

A random sample of 10,000 and 100,000 websites was crawled by the project team using Heritrix in order to simulate a broad crawl. This sample was taken from the list of URLs that was established in task 2.1. For the selective thematic collections, the seed lists created by the State Archives and KBR in task 3.1 were crawled in their entirety.

Task 3.4: Piloting access to the Belgian web archive for scientific research (URF-SID and Ghent-CDH)

Access to the Belgian web archives varies according to the type of use, so there may be significant differences between the general public and the scientific community. To address these differences, several approaches were explored.

A first way to provide access to the web archive collections is based on WARCLight which is an extension of an existing discovery tool called Blacklight. This discovery tool uses Ruby on Rails and Solr to provide a rich interface for full-text and faceted searches. WARCLight helps shed light on the content in web archives (Archives Unleashed (2019)) and helps the user to find what s/he needs. When a resource is found, the application displays some metadata about the resource and its capture. It also provides a link to the replay of the archived webpage.

Another way of accessing a web archive is replaying websites that were captured at a specific point in time. Within the project this was done based on an implementation of the Wayback Machine in Python, called PyWB, that allows search based on URL and timestamp (Webrecorder (2019b)). PyWB also implements the Memento Protocol which enables the user to access HTTP resources based on a specific point in time.

The two previous approaches are intended a priori for all audiences. However, for the scientific community, the plan was to offer other types of access. A first example is to publish the derivative files from the collections on a website that can be used for different types of analysis as discussed above. Another way is to offer researchers access to the WARC files which requires an agreement between the State Archives and KBR and researchers. Another example is to give access to specific collections and datasets, such as the sub-archives, that were mentioned above.

Task 3.5: Evaluation of the pilot Belgian web-archive (MICT and URF-SID)

The captured content as well as the implemented tools were evaluated in an iterative and informal way during the lifetime of the project. This included an exercise to assess the quality and completeness of archived web material. In this context we also focussed on research questions such as (i) ‘What percentage of Belgian history is lost as a result of the lack of a Belgian web-archive?’, (ii) ‘What websites resisted time and are still online?’ and (iii) ‘How much of the Belgian web of the past can be reconstructed through other web-archives or using other ‘web archaeology’-techniques?’. Results

were (amongst others) presented in the conference paper 'Unearthing the Belgian web of the 1990's: a digitised reconstruction' which was presented at 'The Web That Was: Archives, Traces and Reflections', the 3rd RESAW Conference (Amsterdam, 19-21 June 2019). For other evaluation activities the personas created in the Corpus project and outlined in the deliverable 'Le projet Corpus et ses publics potentiels: Une étude prospective sur les besoins et les attentes des futurs usagers' (see https://hal-bnf.archives-ouvertes.fr/hal-01739730/document) were used. As such, these five personas, initially described in order to help with the identification of potential users "(…) l'identification et à la définition des profils des usagers potentiels" (p. 38), were used to assess current access methods to, and analysis tools for, web archives.

Work Package 4: Recommendations for sustainable web-archiving in Belgium (lead: KBR)

The aim of this work package was to make recommendations for a sustainable web archiving service in Belgium and was divided into five tasks.

Task 4.1: Legal considerations concerning access to web-archives (AGR, KBR and CRIDS)

Since all access considerations related to copyright were fully processed and developed in WP2, WP4 focused on the issue of protection of personal data. The report of WP4 contains an analysis of the provisions, rules and principles of personal data protection relevant for the web archiving activities of KBR and the State Archives. The report contains a series of summary boxes designed to reflect the principles, various obligations and possible derogations at the operational level, taking into account the interests of the KBR and the AGR.

First, we compiled a glossary of key data protection terms to facilitate the understanding of subsequent developments. In a second step, we explained the state of the art of personal data protection by detailing the different principles that should guide any data processing within the scope of the GDPR. Thirdly, we analysed the derogation regime provided for in the GDPR for the processing of personal data for archiving purposes in the public interest. In a fourth and final step, we reported on the additional specificities contained in Belgian law for the processing of personal data for archiving purposes in the public interest.

The results of this WP4 report were also valorised as a publication on data processing for archiving purposes in the public interest for the *DPO News* journal.

Task 4.2: Business model development (AGR and KBR)

The business model framework that has been chosen as the basis for this study is the Service Dominant Business Model. (Lüftenegger et al. (2013), Traganos et al. (2015), Turetken et al. (2018)). Most business models are geared towards a manufacturing economy, which does not suit web archiving activities very well. Therefore a model based on the service economy was chosen to better accommodate the context of web archiving.

The Service Dominant Business Model (SDBM) is based on four pillars that each consist of a number of elements. Table III provides an overview of the pillars and key questions in the SDBM. The four pillars are: service, actors, management and cost-benefit. **Service** is about the benefit of the value that is 'co-created by the usage of the good' or the value-in-use. **Actors** are the participants in the business while **management** covers both the management of relationships and service flows between the actors. **Cost-benefit** covers financial costs but also other costs and benefits such as knowledge (Lüftenegger et al. (2013), p. 4, 15-16, 30).

Table III**:** Pillars and key questions in the Service Dominant Business Model

Pillar	**Pillar key question**
Service	What is the value-in-use for the user?
Actors	Who are the actors participating in the business?
Management	How are the actors participating in the business?
Cost-Benefit	How much cost and benefit do the actors incur or obtain within the business?

The Service-Dominant Business Model is visualised by means of a radar (Figure 2). In the centre of the radar is the co-created value-in-use proposition that proposes a solution to a customer's problem or customer's experience. The first layer represents the value proposition that each actor contributes to the co-created value-in-use. The second layer lists the co-production activities of each actor which are essentially activities that each actor performs to achieve the co-creation of value. The third layer, actor cost/benefits, lists the (non-)financial expenses or gains of the co-creation actors. The concentric circles are divided by actor into different pie slices. The focal organisation and the customer are given a specific colour. The focal organisation is usually the organisation or organisations that set up the business model.

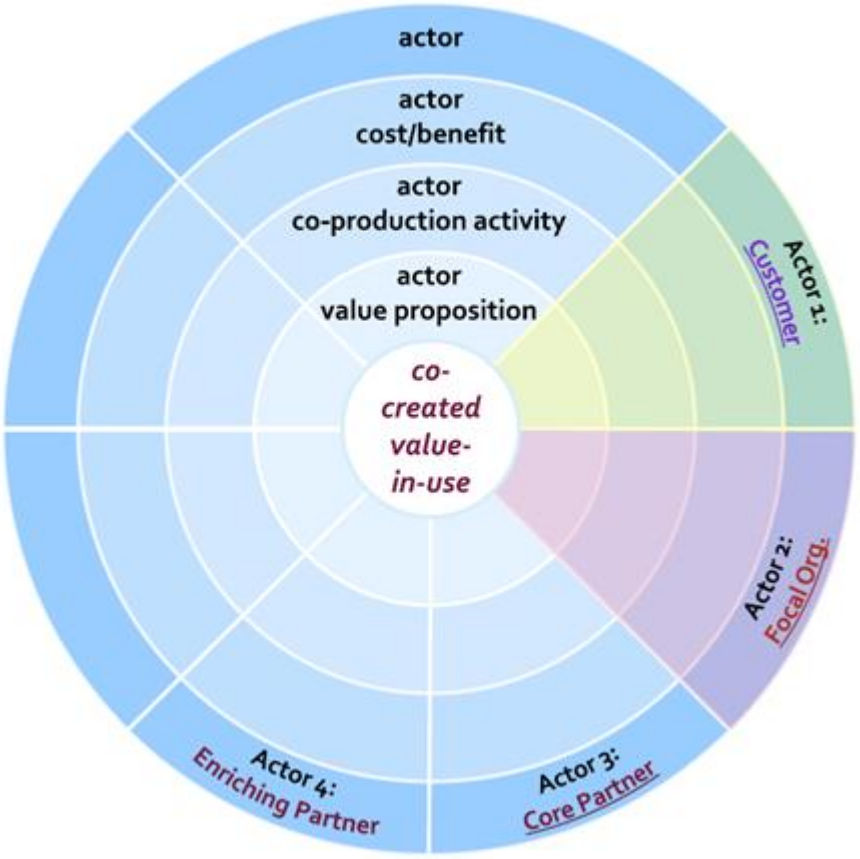


Figure 2: Service Dominant Business Model visualisation (Turetken (2018), p. 17)

KBR and the State Archives were identified as the focal organisations. Other stakeholders in the business model of KBR and the State Archives are: external service providers, government bodies subject to the Law on Archives, other Belgian heritage institutions, the users of the web archive, IIPC (International Internet Preservation Consortium) (IIPC (2019)), RESAW (A Research Infrastructure for the Study of Archived Web Materials) (RESAW (2019)), BELSPO, the regulatory level in Belgium, Belgian society and the registries for Belgian domain names.

Three different scenarios were outlined in the business model: a Belgian web archive managed by KBR and AGR that is managed internally, a web archive that is managed internally but offers web archiving services to other heritage institutions and one in which a third-party service provider manages the Belgian web archive for the AGR and KBR. In order to render the interactions between stakeholders more easily comprehensible, a BPMN (Business Process Model and Notation) mapping was created for each of these three scenarios. A SWOT-analysis was also done for a functional web archive at the State Archives and KBR. These different analyses resulted in the creation of 3 business model radars, one for each of the three different scenarios.

Task 4.3: Legal recommendations for web-archiving (CRIDS)

In terms of recommendations for sustainable web archiving, several actions have been undertaken. First of all, several recommendations have been made to revise and to modify the legal deposit legislation in Belgium. Secondly, an in-depth report about legal considerations concerning access to

web archives was drafted. Thirdly, decision trees have been created for copyright reasoning to be used for web archiving at both the selection and access stage. Fourthly, a list of operational FAQs was developed that can be used by the State Archives and KBR as guidelines for personal data protection in the context of web archiving.

Task 4.4: Technical and functional requirements for an operational web-archiving system for Belgium (URF-SID)

The aim of this task was to outline the technical and functional requirements for an operational web archiving system. This task was achieved in two ways. First, a list of technical and functional requirements was drafted covering the different phases of the OAIS-model with the addition of Selection and Capture. The requirements were divided into four categories according to the MoSCoW model: Must have, Should have, Could have and Won't have. The tools were either selected or developed based on this list of requirements.

Secondly, documentation was provided to enable the deployment of all the tools developed during the project. Particular attention was paid to the documentation of the tools for immediately starting the archiving of the Belgian web, namely the selection tool (internal development), the crawling tool (Heritrix), the replay tool (PyWB) and the search tool (WARCLight). As far as possible, the lessons learned from the different crawls have been integrated into the documentation of the different software.

In addition to the documentation of the tools used, an estimate of the tasks required to set up and maintain archiving of the Belgian web was also provided.

Task 4.5: Definition of procedures (AGR, KBR and CRIDS)

The purpose of this task was to provide procedures and guidelines for users internal and external to KBR and the State Archives. Internal users were interpreted as archivists, librarians and administrative staff while the external stakeholders are web content creators and administrators, end users of the web archives and the general public. These procedures and guidelines have been drafted taking into account the research results obtained from the work packages during the project including the different legal analyses (definition of the Belgian web, legal missions of and legislation related to the State Archives and KBR, considerations concerning access to web archives, legal recommendations, …), the state of the art (i.e. practices identified at other web archiving institutions abroad), the analysis of the user requirements, the results of the pilot, the strategy and business plan. Web archiving procedures and FAQs of other web archiving institutions have also been used as sources of inspiration. In that light the procedures can be considered as an overarching result, incorporating the research results of the entire PROMISE project.

The part for internal users covers the entire web archiving process: selection, collecting and quality control, ingest, storage, data management, preservation planning, access, administration and strategic management. Two specific sections on the user requirements that were studied by means of

a survey undertaken by the University of Ghent and the legal considerations to take into account for a functional web archiving process were also included.

The part for users external to KBR and the State Archives has been formulated as Frequently Asked Questions (FAQs).

Work Package 5: Dissemination, valorisation and coordination (lead: KBR)

Task 5.1 Dissemination activities

Task 5.2 Valorisation activities

Task 5.3 Coordination and management

See Section 5. Dissemination and valorisation for more information on these tasks.

Secondary research results

The research project achieved a number of secondary research results. KBR became a member of IIPC (International Internet Preservation Consortium) which allowed the institution to tap into the international web archiving community. The IIPC provides a lot of support for web archiving institutions worldwide and functions as a platform for information exchange and expertise. It is a very important network for KBR since a lot of inspiration can be drawn from the events they organise and fund. Several members of the PROMISE project team also became members of WARCnet, a project initiated by Niels Brügger that starts in 2020 and aims to bring together researchers and web archiving professionals from all over the world by providing funding for travel expenses and organising conferences. It goes without saying that the PROMISE project facilitated our inclusion in this project. Several members of the PROMISE team also are included in the COST action 'Digital cultures in Europe through web archives' that aims "to structure and develop a transnational interdisciplinary network and platform for researchers who study the archived web and for web archivists". This will also further information exchange and building expertise.

A further outcome that was not specified in the PROMISE project proposal was the creation of a report on the status of web archiving in Belgium. Meetings were set up with representatives of Belgian web archiving institutions that are involved in web archiving: Felixarchief, Vlaams Instituut voor Archivering, Universiteitsbibliotheek Gent, Liberaal Archief, AMSAB - Instituut voor Sociale Geschiedenis, ADVN - Archief voor Nationale Bewegingen, Letterenhuis, KADOC - Documentation and research centre on religion, culture and society, Architectuurarchief Vlaanderen, Archief Gent, Université Catholique de Louvain and the research project 'Catching the digital heritage' (Liberas (2019)) at AMSAB and Liberaal archief. One of the ideas the State Archives and KBR can pursue in the future is to set up a shared catalogue of metadata of all web archive collections in Belgium, similar to what has been set up in The Netherlands (Netwerk Digitaal Erfgoed, (2019)). The fact that contact has been made with these organisations would facilitate accomplishing this task.

The PROMISE project also gave the team members opportunities to learn new skills. Two members of the PROMISE project team had the opportunity to follow an online course in web archives and web archiving at NetLab in Denmark at the beginning of the project, which was an excellent introduction to web archiving from a technical and operational level. (Netlab (2019)) Hands-on experience with tools such as Heritrix, Browsertrix and Brozzler was obtained. The Archives Unleashed Toolkit was tested by a number of team members as well as Gephi, a tool for data visualisation. A number of cataloguers at KBR also got an introduction to web archiving as a new activity for the library and received training on how to select relevant materials and create descriptive metadata. Training material (powerpoint presentations, collection descriptions, templates, ...) was also developed to this end. The ICT department at KBR and the State Archives also followed the project from the sidelines and therefore acquired new insights into web archiving from a technical point of view. The work on a prototype for automated quality assurance and the creation of derivative files that was not included in the original project proposal also required developing new skills.

The PROMISE project also gave the opportunity to employ a number of interns who worked on separate elements. Patricia Blanco for example, a masters' student in Digital Humanities, created three collections for KBR on the representation of the Portuguese, Italian and Spanish communities in Belgium for her master thesis. Amory Hoste, a masters' student in computer sciences at Ghent University, conducted research on techniques to be applied in the project's software prototype, for instance quality assurance and archive derivatives.

The findings of the PROMISE project team also enabled another Belgian research project to draw inspiration for their project proposal: 'Catching the digital heritage'. Furthermore, the findings inspired another research project proposal for the BRAIN.be programme to be submitted, coordinated by KBR. The research proposal centres on the archiving of social media and the information gathered during the PROMISE project was instrumental for setting up this project proposal.

The impact of the research results on science, society and policy

The impact of the research results of the PROMISE project on science are numerous. On the Belgian level, an in-depth study of web archiving practices has never been conducted. Furthermore, one of the strengths of the PROMISE project is the fact that it was centred around an interdisciplinary approach and brought together technical, legal and operational experts and experts in digital humanities, information management, communication studies and user needs. On an international level, the article based on the study of the state of the art that was conducted during the first research phase gave an in-depth and comparative overview of the web archiving practices at different institutions abroad. This provided an important contribution to the scientific literature about web archiving. The importance of the article is also underlined by the large number of downloads. At the beginning of December 2019, the article has been downloaded more than 81.000 times. (Vlassenroot et al. (2019))

The impact on Belgian society will increase the longer and larger the web archive collections grow. As has been argued, without a web archive at the national level, a significant and central part of

contemporary history will become unavailable for future research, leading to a digital dark age. Setting up a Belgian web archive based on the findings of the PROMISE research project will prevent this from happening. This makes the research results very relevant for society in the long run since the main aim is to preserve Belgian heritage. Furthermore, the survey that was undertaken reached different audiences: researchers and academics, the general public and professionals in the heritage sector. This constitutes an important channel of outreach about the project and touched different parts of Belgian society as well as the international web archiving community.

With regards to policy, the State Archives and KBR will present a strategic note to the government based on the findings of the PROMISE project in order to underline the necessity of web archiving as a new mission for both institutions. KBR and the State Archives will also work towards making the necessary changes to the legislation related to their activities based on the legal recommendations formulated during the project.

On a more operational level, the PROMISE project has inspired the creation of a number of policy documents such as a web archiving policy including cost calculation, a business model, FAQ based on the legal recommendations, … The findings of the PROMISE project will be instrumental to the creation of a sustainable web archiving service on the federal level in Belgium.

4. DISSEMINATION AND VALORISATION

4.1 Conferences attended

- Dodging the Memory Hole - saving online news, San Francisco, 15-16 November 2017
- W3C Publishing Summit, San Francisco, 9-10 November 2017
- L'exception au droit d'auteur pour copie privée et la compensation du préjudice qui en résulte dans un environnement dématérialisé : défis et perspectives, Bruxelles, 16 octobre 2017.
- L'intelligence artificielle et le droit (conférence @CRIDS), Namur, 20 octobre 2017.
- Trust (in) the Digital Transition: Management and archiving challenges facing a new legal & standard framework, Brussels, 29 November 2017.
- Seminar Lexalert: GDPR - Zoom sur deux obligations pour le responsable du traitement et pour le sous-traitant : le délégué à la protection des données et le registre des activités de traitement, Webinar, 25 janvier 2018.
- RGPD : et si on s'y mettait ?, Namur, 13 mars 2018.
- EU Copyright, quo vadis? From the EU copyright package to the challenges of Artificial Intelligence, Brussels, 25 May 2018.
- EU Copyright Reform & Why Libraries should care, Webinar, 4 June 2018.
- DARIAH workshop on metadata, Workshop, Brussels, 14-15 May 2018.
- What's in the web archive, Workshop, den Haag, 10 October 2018.
- IIPC Web archiving conference 2018, Wellington (New-Zealand), 12-15 November 2018.
- Three decades @ the crossroads of IP, ICT and Law (30 years of CiTIP/KULeuven), Leuven, 4 October 2019.

4.2 Presentations given

- Chambers, S., Mechant, P., Vandepontseele, S., Isbergue, N. and Depoortere, R. *Towards a national web in a federated country: a Belgian case study*. Presentation at Workshop on National Webs, Aarhus University and the State Library Denmark, 8-9 December 2016.
- Chambers, S., Mechant, P., Teszelszky, K., and Maurer, Y. *Research opportunities for the archived web in the Benelux*. Presentation at the DH Benelux conference 2017, Utrecht, 3-5 July 2017.
- Chambers, S. & Mechant, P., *Aanslagen, Attentats, Terroranschläge: developing a special collection for the academic study of the archived web related to the Brussels terrorist attacks in March*, at RESAW, London (14 – 15 June 2017).
- Michel, A., *Présentation du projet PROMISE lors du Midi du CRIDS*, Namur, 21 November 2017.
- Geeraert, F., Michel, A., Soyez, S. & Vlassenroot, E., *Presentation of the PROMISE project during InterPARES Trust joint European Team and Transnational Team Workshop*, Brussels, 1st December 2017.
- Di Pretoro, E., Geeraert, F., Michel, A. & Vlassenroot, E., *Presentation of the PROMISE project at the BnF*, Paris, 12 December 2017.
- Geeraert, F., *Sélection et accès pour l'archivage du web*, Presented at Midi du CRIDS, 2 February 2018.
- Chambers, S. *Investigating the PROMISE of a Belgian Web Archive. Presentation at Web archiving: best practices for digital cultural heritage,* Jerusalem, 29 April 2018.
- Di Pretoro, E., *Behind the scenes of web archiving. The metadata of harvested websites.* Presentation at Trust and Understanding: the value of metadata in a digitally joined-up world, State Archives of Belgium, 15 May 2018.
- Depoortere, R., Geeraert, F., Mechant, P., Soyez, S. and Vlassenroot, E., *Web archives as sources for researchers.* Presentation at Dag van de Nieuwste Geschiedenis, KU Leuven, 25 May 2018.
- Di Pretoro, E., Geeraert, F., Michel, A. & Vlassenroot, E., *The promise of web archiving in Belgium.* Presentation at Déjeuner de la recherche at the Royal Library of Belgium, 1 June 2018.
- Vlassenroot, E., *Using the archived web for humanities research. Researcher needs for the future Belgian archive, the PROMISE challenge*. Presentation at Digital Humanities Benelux Conference, Amsterdam, 6 June 2018.
- Geeraert, F., *Critical reflections on unlocking web archives for humanities research*. Presentation at Digital Humanities Benelux Conference, Amsterdam, 8 June 2018.
- Isbergue, N., *PROMISE Preserving Online Multiple Information: Towards a Belgian Strategy*. Poster presented at Printemps Scientifique, State Archives of Belgium, 18th June 2018.
- Geeraert, F., *The promise of web archiving in Belgium.* Presentation at InterPARES Trust Symposium, Jerusalem, 19 June 2018.
- Michel, A., *In bytes we trust ! : Digital Act & Archivage électronique*, Presentation at Association des Archivistes Francophones de Belgique study day, Namur, 5 October 2018.

- Geeraert, F., *The PROMISE project: The road to a Belgian Web Archive*, Presentation at DLM-Forum, Vienna, 29 November 2018.
- Soyez, S., *The new Belgian Web Archiving Strategy & Prototype: a PROMISE*, Presentation at DLM-Forum, Bern, 21 May 2019.
- Soyez, S., *Policies: The missing link (including the PROMISE strategy/policy)*, Presentation at Symposium "Policy Matters" ACA@UBC, University of British Columbia (Vancouver), 15 February 2019.
- Di Pretoro, E. Participation in the *Hackathon on Automated Quality Assurance* organized by IIPC at The National and University Library of Iceland, 3-5 April 2019.
- Geeraert, F, Chambers, S. and Mechant, P., *Towards a national web archive in a federated country: a Belgian case study,* Presentation at the IIPC conference, Zagreb, 6 June 2019.
- Geeraert, F. and Soyez, S., *The first steps towards a Belgian web archive*, Presentation at the IIPC conference Zagreb, 6 June 2019.
- Michel, A., *The legal framework for web archiving: focussing on GDPR*, Presentation at the IIPC conference, Zagreb, 7 June 2019.
- Mechant, P. *Presentation of the PROMISE project and SOTA* at Studiedag 'Het web gearchiveerd?', Nederlands Instituut voor Beeld en Geluid (Hilversum), 11th November 2018: https://www.netwerkdigitaalerfgoed.nl/events/studiedag-het-web-gearchiveerd/.
- Chambers, S., Mechant, P. et al. Workshop '*Hacking the news*' at the third biennial RESAW (Research Infrastructure for the Study of Archived Web Materials) conference 'The web that was: archives, traces, reflections', University of Amsterdam, 18th June 2019: http://thewebthatwas.net/program/.
- Chambers, S., Geeraert, F., Vlassenroot, E., Mechant, P., Haesendonck, G. and Di Pretoro, E., Presentation '*Unearthing the Belgian web of the 1990's: a digitised reconstruction*' at the third biennial RESAW (Research Infrastructure for the Study of Archived Web Materials) conference 'The web that was: archives, traces, reflections', University of Amsterdam, 19th June 2019: http://thewebthatwas.net/program/.
- Di Pretoro, E., Michel, A and Geeraert, F. Presentation '*The road to a Belgian federal web archive*' at Informatie aan Zee, Oostende, 19 September 2019.
- Soyez, S. Presentation '*Stratégie archivage du web belge*' at Dag van de bibliothecaris, Brussels, 5 December 2019.

4.3 Organisation of concluding colloquium 'Saving the Belgian web: the promise of a Belgian web archive' on 18 October 2019

4.3.1 Programme

Morning

8.30-9.00:	**Registration**
9.00-9.15:	**Welcome** Sara LAMMENS (KBR), Dr Karel VELLE (State Archives)
9.15-9.45:	**National web archives: the land of promise for researchers** **Keynote** by Prof. Dr Niels BRÜGGER (Aarhus University)
9.45-10.45:	**Introducing the PROMISE project** Speakers: Sophie VANDEPONTSEELE (KBR); Friedel GEERAERT (KBR/State Archives); Dr Peter MECHANT (UGhent); Alejandra MICHEL (CRIDS - UNamur)
10.45-11.15:	Coffee break
11.15-11.50:	**Towards a Belgian strategy** Speakers: Sébastien SOYEZ (State Archives); Dr Rolande DEPOORTERE (State Archives); Sophie VANDEPONTSEELE (KBR)
11.50-12.20:	**Technological aspects** Speaker: Gerald HAESENDONCK (UGhent)
12.20-13.30:	Lunch break

Afternoon

13.30-14.45:	**Access to collections for the broad public (panel)** Speakers: Pierre DE MÛELENAERE (ONLIT Editions); Philippe NOTHOMB (Rossel group); Dr Tine VEKEMANS (AMSAB); Prof. Dr Cécile DE TERWANGNE (CRIDS – Unamur);Sophie VANDEPONTSEELE (KBR); Dr Rolande DEPOORTERE (State Archives)
14.45-15.15:	Coffee break
15.15-16.30:	**Using web archives for researchers** Speakers: Sally CHAMBERS (GhentCDH); Eveline VLASSENROOT (UGhent); Prof. Dr Valérie SCHAFER (University of Luxembourg); Jesse DE VOS (Beeld & Geluid); Patricia BLANCO
16.30-16:45:	**Synthesis of the two panels** by Prof. Dr Valérie SCHAFER (University of Luxembourg)

4.3.2 Short report

On 18 October 2019, the colloquium 'Saving the Web: the Promise of a Belgian Web Archive' took place at KBR. The colloquium was organised by the researchers of the PROMISE research project which aims to develop a federal strategy for the preservation of the Belgian web. Web archiving is clearly a 'hot topic' because the event attracted 107 participants from Belgium and abroad.

The need for a web archive

The Directors of KBR and the State Archives stressed the need for structural funding for the development and maintenance of a web archive at the federal level within both institutions. This was also confirmed by the keynote speaker, Prof. Niels Brügger: we must at all costs avoid losing another 25 years of Belgian web history. Without a national web archive it is not possible to perform (trans)national historical analysis based on the web. The best conditions for this:

- one archive of the national web, based on a web archiving strategy that is as complete as possible
- good access possibilities, including data extraction by or for researchers so that they can work with the data themselves on their own computer

Best practices

After the inspiring keynote, the researchers gave an overview of the research results of the PROMISE project. Among other things, the best practices within web archiving were presented, as well as the results of the legal analysis, the strategy and various scenarios developed by the State Archives and KBR, the technical specifications of the web archiving process and the results of the survey on user needs within web archives.

Access to the web archive

In the afternoon, the focus was on access, both for the general public and for research. Representatives of heritage institutions, the publishing world and the press took part in the panel discussion on access for the general public. The discussion revealed that copyright legislation severely restricts access to web archives. According to the letter of the law, in most cases web archives can only be made available in the reading room without the explicit permission of the right holders, which of course constitutes a serious obstacle.

The need to highlight not only the costs but also the many advantages of web archiving was also underlined. For example, the State Archives and KBR could develop many additional services on the basis of the data in the web archive, which constitute a real added value for society and research.

Research in web archives

In the session on access to the web archive for research, Jesse de Vos of Sound and Vision presented the Dutch initiative 'National Register Web Archives'. This is a great project that can serve as an inspiration for Belgium. In her presentation, Prof. Valérie Schafer reflected on the when, what and

why of 30 years of world wide web and stressed why web archives are so important to researchers. Patricia Blanco explained how during her internship at KBR she started working with the web archive collections of KBR for her masters' thesis 'Saving the Belgian web: web archiving practices, research opportunities and limitations'.

Questions from the audience

During the Q&A sessions, the audience showed interest in archiving the .eu domain. The web archive in Portugal, Arquivo.pt, has set up a pilot project to archive this top level domain and the EU Publications Office archives all *.europa.eu* domain names.

The concrete agenda for the further development of the Belgian web archive was also discussed. This largely depends on the new government as additional funds are needed to develop the web archive.

Other points that were raised were the need to focus on users and to inform and involve them sufficiently, the advantages of using the WARC file format and the arguments that can be decisive to convince the political decision makers of the need for web archiving and therefore also the need for structural funding.

The colloquium resulted in cross-fertilisation between the various professionals who are involved in web archiving or the study of archived websites. The PROMISE project team is in any case entering the final months of the project full of new inspiration.

This report and the slides of the presentations given during the colloquium can be downloaded via this link: https://www.kbr.be/en/colloquium-saving-the-web/.

5. PUBLICATIONS

5.1 Peer review publications

5.1.1 Chapters

Chambers, S., Mechant, P. & Geeraert F. (2019). Towards a National Web in a Federated Country: a Belgian Case Study. In N. Brügger & D. Laursen (Eds.), *The Historical Web and Digital Humanities: the Case of National Web Domains*, Routledge. https://biblio.ugent.be/publication/8610092

5.1.2 Articles

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1(1); DOI 10.1007/s42803-019-00007-7; https://link.springer.com/article/10.1007/s42803-019-00007-7.

Di Pretoro, E. (2019). IIPC -Hackathon sur l'automatisation des contrôles qualité. *Cahiers de la documentation*, 2019/2.

Di Pretoro, E., Geeraert, F. and Soyez, S. (2019). Behind the Scenes of Web Archiving: Metadata of Harvested Websites. *ABB: Archives et Bibliothèques de Belgique -Archief-en Bibliotheekwezen in België*

in Trust and Understanding: the value of metadata in a digitally joined-up world. Eds. R. Depoortere, T. Gheldof, D. Styven and J. Van Der Eycken, 106, pp.63-74. https://hal.archives-ouvertes.fr/hal-02124714/document

A. Michel, "Les traitements de données à des fins archivistiques dans l'intérêt public", *DPO News*, 2019/5, pp. 6 à 9.

5.1.3 Others

Mechant, P. (2019). Web 25 Book review. *International Journal of Digital Humanities*, 1(1).; https://link.springer.com/article/10.1007/s42803-019-00010-y.

5.2 Publications (no peer review)

5.2.1 Articles

Geeraert, F. (2018-2019). De weg naar een federaal webarchief. *Science Connection*, 59, p. 32-34. http://www.belspo.be/belspo/organisation/publ_science_en.stm

5.2.2 Conference posters

Geeraert, F. (2018). *PROMISE Preserving Online Multiple Information: Towards a Belgian Strategy*. [Poster].

Michel, A. (2018). *Web archiving: need for a Belgian federal strategy (about PROMISE project).* [Poster].

5.2.3 Conference abstracts

Depoortere, R., Geeraert, F., Mechant, P., Soyez, S. and Vlassenroot, E., *Web archives as sources for researchers*. Dag van de Nieuwste Geschiedenis 2018, KULeuven - Leuven. [abstract]

Geeraert, F., Michel, A., & Vlassenroot, E., *Critical reflections on unlocking web archives for humanities research*. DH Benelux 2018, Amsterdam. [abstract] Available online at http://2018.dhbenelux.org/wp-content/uploads/sites/8/2018/05/Friedel-Geeraert-Alejandra-Michel-Eveline-Vlassenroot_Critical-reflections-on-unlocking-web-archives-for-humanities-research_DHBenelux2018.pdf.

Geeraert, F. & Soyez, S., *The promise of web archiving in Belgium*. Interpares Trust 2018, Jerusalem. [abstract]

Geeraert, F. & Soyez, S., *The road to a Belgian federal web archive*. DLM Forum 2018, Vienna. [abstract]

Geeraert, F., *The first steps towards a Belgian web archive on the federal level.* IIPC General Assembly and Web Archiving Conference 2019, Zagreb. [abstract] Available online at http://netpreserve.org/ga2019/wp-content/uploads/2019/07/IIPCWAC2019-FRIEDEL_GEERAERT__SEBASTIEN_SOYEZ-The_first_steps_towards_a_Belgian_web_archive-a_federal_strategy.pdf

Chambers, S., Geeraert, F., Vlassenroot, E., Mechant, P., Haesendonck, G., Di Pretoro, E. (2019). *Unearthing the Belgian web of the 1990's: a digitised reconstruction*. RESAW 2019, Amsterdam. [abstract] Available online at https://easychair.org/smart-program/RESAW19/2019-06-19.html#talk:89175

Di Pretoro, E., Michel, A. & Geeraert, F. (2019). *The Road to a Belgian federal web archive*. Informatie aan Zee 2019, Oostende. [abstract] Available online at https://www.vvbad.be/activiteiten/informatie-aan-zee-2019/road-belgian-federal-web-archive-de-weg-naar-een-federaal

Chambers, S., Geeraert, F., Mechant, P. and Vlassenroot, E. (2019). *Piloting access to the Belgian web-archive for scientific research: a methodological exploration.* Engaging with Web Archives conference - Opportunities, Challenges and Potentialities 2020, Maynooth. [abstract]

5.2.3 Others

Geeraert, F. and Willaert, T., (2019). [interview] *Through the black hole of information - Friedel Geeraert on building a Belgian Web Archive*, March 2019, Digital scholarship blog, https://www.digitalscholarship.be/.

Geeraert, F. (2020). Verslag studiedag Saving the web: the promise of a Belgian web archive. *META Tijdschrift voor bibliotheek en archief.* - in press.

6. ACKNOWLEDGEMENTS

First of all, we would like to thank the members of the scientific follow-up committee for providing us with interesting ideas, feedback and guidance throughout the project.

- Filip Boudrez, Archivist – Stadsarchief Antwerpen
- Els Breedstraet, Coordinator web preservation – Publications Office of the European Union
- Aurore François, Director of Archives, UCL
- Seth Van Hooland, Professor - ULB
- Rony Vissers, coordinator Packed vzw
- Peter Webster, Managing director, Webster Research and Consulting Ltd.

We would also like to thank the representatives of web archiving initiatives in Belgium and abroad whom we interviewed or to whom responded to questionnaires for establishing a state of the art. This was a very instructive phase for the entire research team and it also allowed us to start to build a network within the (inter)national web archiving community:

- National Library of The Netherlands: Kees Teszelszky (Researcher web archiving, Digital Preservation Department)
- National Archive of The Netherlands: Antal Posthumus (Adviser recordkeeping, Directie Infrastructuur & Advies) and Jeroen van Luin (Acquisition and Maintenance of Digital Archives)

- National Library of France (BnF): Pascal Tanésie (Assistant to the head of the department of digital legal deposit), Sara Aubry (Web Archiving Project Manager, IT department) and Bert Wendland (IT Department)
- National Library of Luxembourg: Yves Maurer (Webarchiving Technical Manager) and Ben Els (Digital Curator)
- The Royal Danish Library: Jakob Moesgaard (Specialkonsulent, Department of Digital Legal Deposit and Preservation) and Tue Hejlskov Larsen (IT analyst)
- The UK National Archives: Tom Storrar (Head of Web Archiving) and Claire Newing (Web Archivist)
- The British Library: Jason Webber (Web Archiving Engagement and Liaison Manager)
- Arquivo.pt.: Daniel Gomes (Head of Arquivo.pt., the Portuguese web-archive, Advanced Services Department)
- National Library of Ireland (NLI): Maria Ryan (Web Archivist)
- National Library of Switzerland: Barbara Signori (Head of e-Helvetica)
- Bibliothèque et Archives nationales de Québec: Carole Gagné (Direction du dépôt légal et de la conservation des collections patrimoniales)
- Library and Archives Canada: Nathalie Villeneuve (Director digital preservation and migration).and Tom Smyth (Manager digital integration)
- Felixarchief, Antwerp: Inge Schoups, Filip Boudrez en Giovanna Visini
- Vlaams Instituut voor Archivering (VIAA): Nico Verplancke en Matthias Priem
- Universiteitsbibliotheek Gent: Paul Bastijns en Patrick Hochstenbach
- Liberaal Archief: Jeroen Buysse
- AMSAB - Instituut voor Sociale Geschiedenis: Maarten Savels
- ADVN - Archief voor Nationale Bewegingen: Tom Cobbaert and Sophie Bossaert
- Letterenhuis: Isabelle van Ongeval
- KADOC - Documentation and research centre on religion, culture and society: Katrien Weyns
- Architectuurarchief Vlaanderen: Wim Lowet
- Archief Gent: Pieter-Jan Lachaert, Tom Haeck, Steven Staelens and Lien Ceûppens
- Université Catholique de Louvain: Aurore François
- Research project ‘Catching the digital heritage’ (collaboration between AMSAB and Liberaal Archief): Jeroen Buysse, Jeroen Fernandez-Alonso and Tine Vekemans

We would also like to thank Georges Jamart from BELSPO for his follow-up and guidance throughout the entire project.

The speakers who participated in the concluding colloquium ‘Saving the Belgian web: the promise of a Belgian web archive’ also deserve a special thank you, especially since they are advocates for (the study of) web archiving:

- Prof. Dr Niels Brügger (Aarhus University)
- Pierre De Mûelenaere (ONLIT Editions)

- Philippe Nothomb (Rossel Group)
- Dr Tine Vekemans (AMSAB)
- Prof. Dr Valérie Schafer (University of Luxembourg)
- Jesse de Vos (Beeld & Geluid)
- Patricia Blanco (intern at KBR)

REFERENCES

Agence nationale de la Recherche. (s.d.). *Web heritage and history in the 90s – WEB90.* Available online at https://anr.fr/Project-ANR-14-CE29-0012. Last accessed 21 November 2019.

Archives Unleashed. (2019). *Warclight*. Available online at https://github.com/archivesunleashed/warclight. Last accessed on 21 November 2019.

Archives Unleashed Cloud. (2019). *Using the Archives Unleashed Cloud Derivative Files*. Available online at: https://cloud.archivesunleashed.org/derivatives. Last accessed on 22 November 2019.

ASAP (Archives sauvegarde attentats Paris). (s.d.). *A propos du projet ASAP*. Available online at https://asap.hypotheses.org/a-propos. Last accessed 21 November 2019.

Ben-David, Anat and Matamoros-Fernández, Ariadna. (2016). Hate speech and covert discrimination on social media: monitoring the Facebook pages of extreme-right political parties in Spain. *International Journal of communication*. 10 (2016), pp. 1167-1193.

Brügger, Niels. (2014). *Probing a nation's web domain – the historical development of the Danish web*. Available online at https://pure.au.dk/portal/en/projects/probing-a-nations-web-domain--the-historical-development-of-the-danish-web(bfc8df2d-c452-43df-b2ea-fd0d06fcca42).html. Last accessed 21 November 2019.

Common Crawl. (2019). *Common Crawl*. Available online at https://commoncrawl.org/. Last accessed on 21 November 2019.

Costa, M. & Silva, M. (2010). *Understanding the information needs of web archive users.* In Proceedings of the 10th International Web Archiving Workshop (pp. 9-16).

Costa, M. & Silva, M. (2011). *Characterizing search behavior in web archives.* In Proceedings of the 1st International Temporal Web Analytics Workshop.

Costea, M.-D. (2018). *Report on the scholarly use of web archives.* Aarhus: Netlab. Available online at http://netlab.dk/wp-content/uploads/2018/02/Costea_Report_on_the_Scholarly_Use_of_Web_Archives.pdf.

DNS Belgium. (2019). *DNS Belgium. Registry for .be, .brussels and .vlaanderen*. Available online at https://www.dnsbelgium.be/en. Last accessed on 21 November 2019.

Dooley, Jackie & Bowers, Kate. (2018). *Descriptive metadata for web archiving: recommendations of the OCLC Research Library Partnership Web Archiving Metadata Working Group*. Available online at: https://www.oclc.org/content/research/publications/2018/oclcresearch-descriptive-metadata/recommendations.html. Last accessed on 21 November 2019.

Elshobaky, Sara and Eldakar, Youssef. (2019). Identifying Egyptian Arabic websites using machine learning during web crawl. *Conference paper: IIPC General Assembly and Web Archiving Conference*.

Gebeil, Sophie. (2015). *La fabrique numérique des mémoires de l'immigration maghrébine sur le web français (1999-2014).* PhD thesis. Ecole Doctorale Espaces, Cultures Sociétés (Aix-en-Provence).

Gephi. (2019). *The Open Graph Viz Platform*. Available online at https://gephi.org/. Last accessed on 22 November 2019.

Huc-Hepher, Saskia. (2016). The material dynamics of a London-French blog: a multimodal reading of migrant habitus. *Modern languages open.* DOI: http://doi.org/10.3828/mlo.v0i0.91.

International Internet Preservation Consortium. (2019). *International Internet Preservation Consortium.* Available online at http://netpreserve.org/. Last accessed on 25 November 2019.

Internet Archive. (2019a). *Heritrix*. Available online at https://github.com/internetarchive/heritrix3/wiki. Last accessed on 21 November 2019.

Internet Archive. (2019b). *Brozzler*. Available online at https://github.com/internetarchive/brozzler. Last accessed on 21 November 2019.

ISO. (2017). *ISO 28500:2017. Information and documentation — WARC file format.* Available online at https://www.iso.org/standard/68004.html. Last accessed on 21 November 2019.

ISO. (2018). *ISO 14721:2012*. *Space data and information transfer systems — Open archival information system (OAIS) — Reference model.* Available online at https://www.iso.org/standard/57284.html. Last accessed on 21 November 2019.

Liberas. (2019). *Catching the digital heritage.* Available online at https://www.liberas.eu/catching-the-digital-heritage/. Last accessed on 9 December 2019.

Lüftenegger, Egon, Comuzzi, Marco, Grefen, Paul & Weisleder, Caren (2013). *The service dominant business model: a service focused conceptualization*. (BETA publicatie: working papers; Vol. 402). Eindhoven. Technische Universiteit Eindhoven.

Maynard, Diana, Dupplaw, David and Hare, Jonathon. (2013). *Multimodal sentiment analysis of social media*. Conference paper: BCS SGAI Workshop on Social Media Analysis.

McTavish, Sarah. (s.d.). *Network graphing archived websites with Gephi*. Available online at https://cloud.archivesunleashed.org/derivatives/gephi. Last accessed on 21 November 2019.

Milligan, Ian. (2015). *Finding community in the ruins of GeoCities: Distantly reading a web archive. UWSpace.* Available online at http://hdl.handle.net/10012/11650.

Milligan, Ian (2016). Lost in the infinite archive: the Promise and Pitfalls of Web Archives. *International Journal of Humanities and Arts Computing.* 10(1). p. 79-80.

Netlab. (2019). *Netlab*. Available online at http://www.netlab.dk/. Last accessed on 9 December 2019.

Netwerk Digitaal Erfgoed (2019). *Nationaal Register Webarchieven*. Available online at https://www.registerwebarchieven.nl/. Last accessed on 25 November 2019.

RESAW. (2019). *RESAW - A Research Infrastructure for the Study of Archived Web Materials*. Available online at https://resaw.eu/. Last accessed on 25 November 2019.

Schroeder, Ralph, & Brügger, Niels (2017). Introduction: The web as history. In N. Brügger & R. Schroeder (Eds.), *The web as history. Using web archives to understand the past and present* (pp. 1–19). London: UCL Press.

Traganos, Konstantinos, Grefen, Paul, den Hollander, Aafke, Türetken, Oktay & Eshuis, Rik (2015). *Business model prototyping for intelligent transport systems: a service-dominant approach. Case study for Praktijkproef Amsterdam Fase 2 deelproject Zuidoost*.

Türetken, Oktay, Grefen, Paul, Gilsing, Rick, Ege Adali, O. (2018). Service-Dominant Business Model Design for Digital Innovation in Smart Mobility. *Business and Information Systems Engineering*.

Vlassenroot, E., Chambers, S., Di Pretoro, E., Geeraert, F., Haesendonck, G., Michel, A., Mechant, P. (2019). Web archives as a data resource for digital scholars. *International Journal of Digital Humanities*, 1(1); DOI 10.1007/s42803-019-00007-7; https://link.springer.com/article/10.1007/s42803-019-00007-7.

Weber, M. S. (2017). The tumultuous history of news on the web. In N. Brugger & R. Schroeder (Eds.), *The web as history. Using web Archives to understand the past and the present* (pp. 83–100). London: UCL Press.

Webrecorder. (2019a). *Browsertrix*. Available online at https://github.com/webrecorder/browsertrix. Last accessed on 21 November 2019.

Webrecorder. (2019b). *PyWB.* Available online at https://github.com/webrecorder/pywb. Last accessed on 21 November 2019.